An Overview of Catastrophic AI Risks

Year: 2023 | Citations: 235 | Authors: Dan Hendrycks, Mantas Mazeika, Thomas Woodside

---

Abstract

Rapid advancements in artificial intelligence (AI) have sparked growing concerns among experts, policymakers, and world leaders regarding the potential for increasingly advanced AI systems to pose catastrophic risks. Although numerous risks have been detailed separately, there is a pressing need for a systematic discussion and illustration of the potential dangers to better inform efforts to mitigate them. This paper provides an overview of the main sources of catastrophic AI risks, which we organize into four categories: malicious use, in which individuals or groups intentionally use AIs to cause harm; AI race, in which competitive environments compel actors to deploy unsafe AIs or cede control to AIs; organizational risks, highlighting how human factors and complex systems can increase the chances of catastrophic accidents; and rogue AIs, describing the inherent difficulty in controlling agents far more intelligent than humans. For each category of risk, we describe specific hazards, present illustrative stories, envision ideal scenarios, and propose practical suggestions for mitigating these dangers. Our goal is to foster a comprehensive understanding of these risks and inspire collective and proactive efforts to ensure that AIs are developed and deployed in a safe manner. Ultimately, we hope this will allow us to realize the benefits of this powerful technology while minimizing the potential for catastrophic outcomes.